CS626 : Natural Language Processing/Speech, NLP and the Web

Lecture 30:
Phonology, syllables; introduce transliteration

Pushpak Bhattacharyya
CSE Dept.
IIT Bombay

1st Nov, 2012

Phonology: Syllables

Basic of syllables

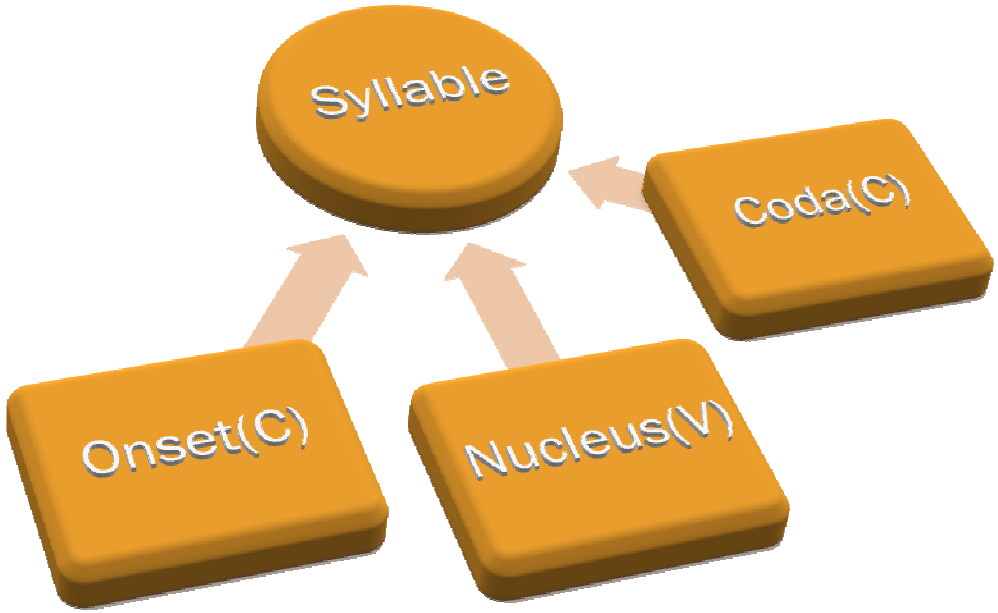
"Syllable is a unit of spoken language consisting of a single uninterrupted sound formed generally by a Vowel and preceded or followed by one or more consonants."

- Vowels are the heart of a syllable (Most Sonorous Element) (*svayam raajate iti svaraH)*
- Consonants act as sounds attached to vowels.

Syllable structure

- A syllable consists of 3 major parts:-
  - Onset (C)
  - Nucleus (V)
  - Coda (C)
- Vowels sit in the Nucleus of a syllable
- Consonants may get attached as Onset or Coda.
- Basic structure - CV

Possible syllable structures



- The Nucleus is always present
- Onset and Coda may be absent
- Possible structures
  - V
  - CV
  - VC
  - CVC

syllable theories

- Prominence Theory
  - E.g. *entertaining* /entəteɪnɪŋ/
  - The peaks of prominence: vowels /e ə eɪ ɪ/
  - Number of syllables: 4
- Chest Pulse Theory
  - Based on muscular activities
- Sonority Theory
  - Based on relative soundness of segment within words

Introduction to sonority theory

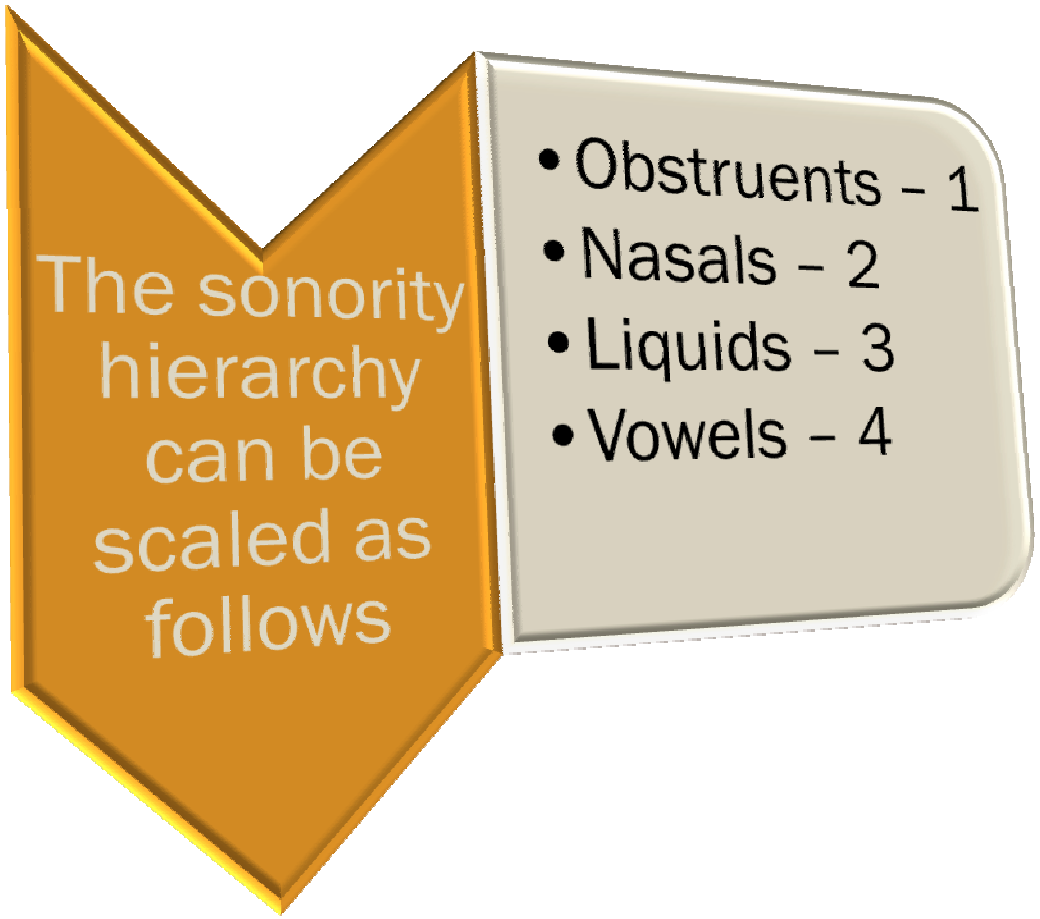
"The Sonority of a sound is its loudness relative to other sounds with the same length, stress and speech."

- Some sounds are more sonorous
- Words in a language can be divided into syllables
- Sonority theory distinguishes syllables on the basis of sounds.

Sonority hierarchy

- Defined on the basis of amount of sound associated
- The sonority hierarchy is as follows:-
  - Vowels (a, e, i, o, u)
  - Liquids (y, r, l, v)
  - Nasals (n, m)
  - Fricatives (s, z, f,…..sh, th etc.)
  - Affricates (ch, j)
  - Stops (b, d, g, p, t, k)

Sonority scale

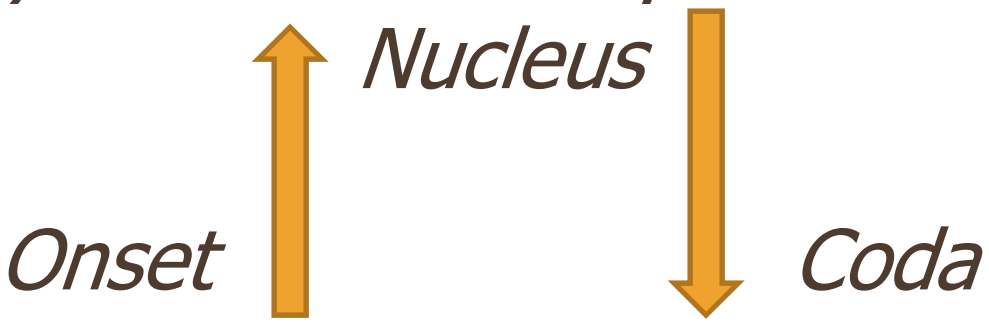


- Obstruents can be further classified into:-
  - Fricatives
  - Affricates
  - Stops

Sonority theory & syllables

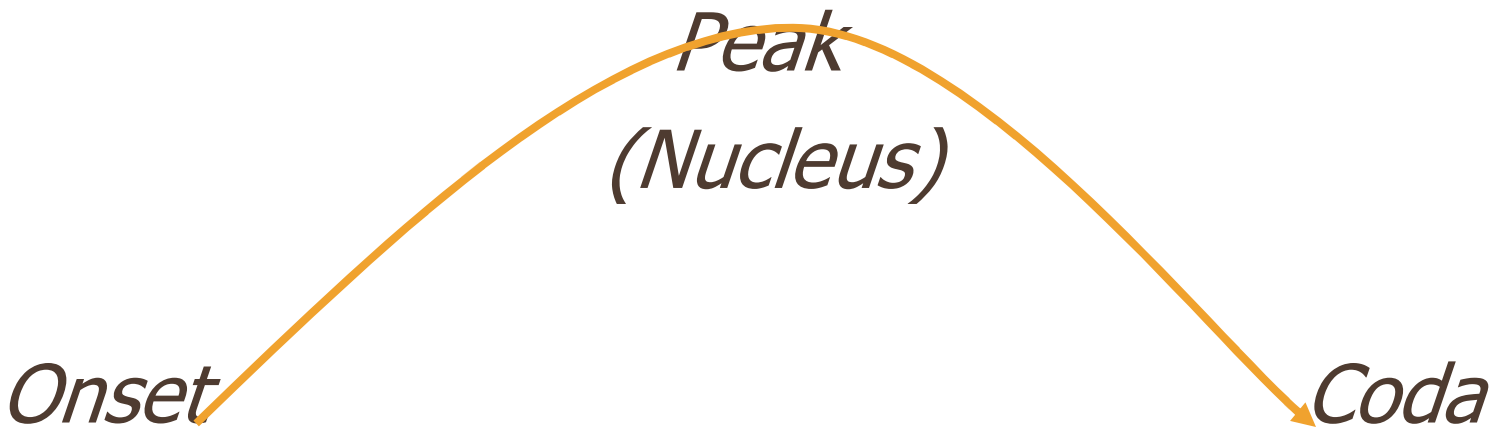
"A Syllable is a cluster of sonority, defined by a sonority peak acting as a structural magnet to the surrounding lower sonority elements."

- Represented as waves of sonority or *Sonority Profile* of that syllable



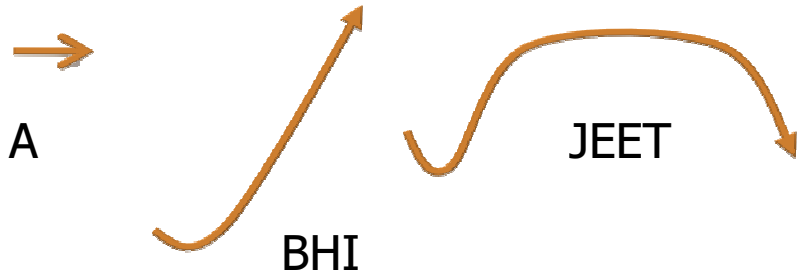
Sonority sequencing principle

"The Sonority Profile of a syllable must rise until its Peak(Nucleus), and then fall."



examples

- ABHIJEET



Profile-1

Profile-2

Maximal onset principle

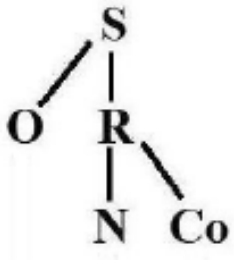
"The Intervocalic consonants are maximally assigned to the Onsets of syllables in conformity with Universal and Language-Specific Conditions."

- Determines underlying syllable division
- Example
  - DIPLOMA

    DIP LO MA & DI PLO MA

Syllable Structure: a more detailed look

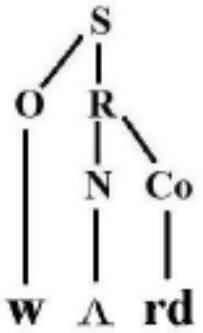
- Count of no. of syllables in a word is roughly/intuitively the no. of vocalic segments in a word.
- Thus, presence of a vowel is an obligatory element in the structure of a syllable. This vowel is called "*nucleus".*
- Basic Configuration: **(C)V(C)**.
- Part of syllable preceding the nucleus is called the *onset*.
- Elements coming after the nucleus are called the *coda*.
- Nucleus and coda together are referred to as the *rhyme*.



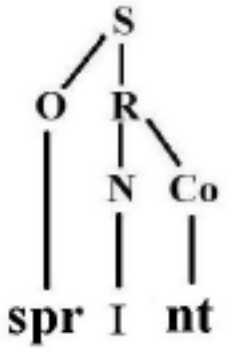
S ≡ Syllable, O ≡ Onset
R ≡ Rhyme, N ≡ Nucleus
Co ≡ Coda

Syllable Structure: Examples

- **‘word’**



- **‘sprint’**

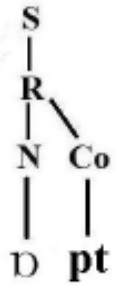


Syllable Structure: Examples

- **‘may’**

← No Coda.

- **‘opt’**



← No Onset.

- **‘air’**

← No Coda, No Onset.

Syllable Structure

- *Open Syllable: ends in vowel*
- *Closed syllable: ends in consonant or consonant cluster*

- *Light Syllable*: A syllable which is open and ends in a short vowel
  - General Description – CV.
  - Example, ‘air’.
- *Heavy Syllable*: Closed syllables or syllables ending in diphthong
  - Example: ‘opt’
  - Example, ‘may’

Syllabification: Determining Syllable Boundaries

- Given a string of syllables (word), what is the coda of one and the onset of another?

- In a sequence such as VCV, where V is any vowel and C is any consonant, is the medial C the coda of the first syllable (VC.V) or the onset of the second syllable (V.CV)?

- To determine the correct groupings, there are some rules, two of them being the most important and significant:
  - Maximal Onset Principle,
  - Sonority Hierarchy

Constraints: Phonotactics

- **Phonotactics**
  - Determines possible comb. of onsets and codas which can occur.
  - Deals with restriction on the permissible comb. Of phonemes.
  - Defines permissible syllable structure, consonant clusters and vowel sequence by means of phonotactical constraints.
- In general, rules operate around the sonority hierarchy.
- Fricative /s/ is lower on the sonority hierarchy than the lateral /l/, so the combination /sl/ is permitted in onsets and /ls/ is permitted in codas. Opposite is not allowed.
- Thus, '*slips*' and '*pulse*' are possible English words.
- '*lsips*' and '*pusl*' are not possible.

Constraints on Onsets

- One-consonant: Only /ŋ/ can't be distributed in syllable-initial position.
- Two-consonant: We refer to the scale of sonority.
  - Sequence 'rn' is ruled out since there is a decrease of sonority.
  - *Minimal Sonority Distance*: Distance in sonority between the first and the second element in the onset must be of at least 2 degrees.
  - Thus, on the basis of *Sonority Hierarchy* and *Minimal Sonority Distance*, only a limited no. of possible two-consonant clusters.
- Three-consonant:
  - Restricted to licensed two-consonant onsets preceded by /s/.
  - Also, /s/ can only be followed by a voiceless sound.
  - Therefore, only /spl/, /spr/, /str/, /skr/, /spj/, /stj/, /skj/, /skw/, /skl/, /smj/ will be allowed. (splinter, spray, strong etc.)
  - While /sbl/, /sbr/, /sdr/, /sgr/, /sθr/ will be ruled out.

Constraints on Onsets

Plosive plus approximant other than /j/	/pl/, /bl/, /kl/, /gl/, /pr/, /br/, /tr/, /dr/, /kr/, /gr/, /tw/, /dw/, /gw/, /kw/	play, blood, clean, glove, prize, bring, tree, drink, crowd, green, twin, dwarf, language, quick
Fricative plus approximant other than /j/	/fl/, /sl/, /fr/, /θr/, /ʃr/, /sw/, /θw/	floor, sleep, friend, three, shrimp, swing, thwart
Consonant plus /j/	/pj/, /bj/, /tj/, /dj/, /kj/, /gj/, /mj/, /nj/, /fj/, /vj/, /θj/, /sj/, /zj/, /hj/, /lj/	pure, beautiful, tube, during, cute, argue, music, new, few, view, thurifer, suit, zeus, huge, lurid
/s/ plus plosive	/sp/, /st/, /sk/	speak, stop, skill
/s/ plus nasal	/sm/, /sn/	smile, snow
/s/ plus fricative	/sf/	sphere

Possible 2-consonant clusters in an Onset

Constraints on Coda

The single consonant phonemes except /h/, /w/, /j/ and /r/ (in some cases)	
Lateral approximant + plosive: /lp/, /lb/, /lt/, /ld/, /lk/	help, bulb, belt, hold, milk
In rhotic varieties, /r/ + plosive: /rp/, /rb/, /rt/, /rd/, /rk/, /rg/	harp, orb, fort, beard, mark, morgue
Lateral approximant + fricative or affricate: /lf/, /lv/, /lθ/, /ls/, /lʃ/, /ltʃ/, /ldʒ/	golf, solve, wealth, else, Welsh, belch, indulge
In rhotic varieties, /r/ + fricative or affricate: /rf/, /rv/, /rθ/ /rs/, /rʃ/, /rtʃ/, /rdʒ/	dwarf, carve, north, force, marsh, arch, large
Lateral approximant + nasal: /lm/, /ln/	film, kiln
In rhotic varieties, /r/ + nasal or lateral: /rm/, /rn/, /rl/	arm, born, snarl
Nasal + homorganic plosive: /mp/, /nt/, /nd/, /ŋk/	jump, tent, end, pink

Constraints on Coda

Nasal + fricative or affricate: /mf/, /mθ/ in non-rhotic varieties, /nθ/, /ns/, /nz/, /nʧ/, /nʤ/, /ŋθ/ in some varieties	triumph, warmth, month, prince, bronze, lunch, lounge, length
Voiceless fricative + voiceless plosive: /ft/, /sp/, /st/, /sk/	left, crisp, lost, ask
Two voiceless fricatives: /fθ/	fifth
Two voiceless plosives: /pt/, /kt/	opt, act
Plosive + voiceless fricative: /pθ/, /ps/, /tθ/, /ts/, /dθ/, /dz/, /ks/	depth, lapse, eighth, klutz, width, adze, box
Lateral approximant + two consonants: /lpt/, /lfθ/, /lts/, /lst/, /lkt/, /lks/	sculpt, twelfth, waltz, whilst, mulct, calx
In rhotic varieties, /r/ + two consonants: /rmθ/, /rpt/, /rps/, /rts/, /rst/, /rkt/	warmth, excerpt, corpse, quartz, horst, infarct
Nasal + homorganic plosive + plosive or fricative: /mpt/, /mps/, /ndθ/, /ŋkt/, /ŋks/, /ŋkθ/ in some varieties	prompt, glimpse, thousandth, distinct, jinx, length
Three obstruents: /ksθ/, /kst/	sixth, next

Other Constraints

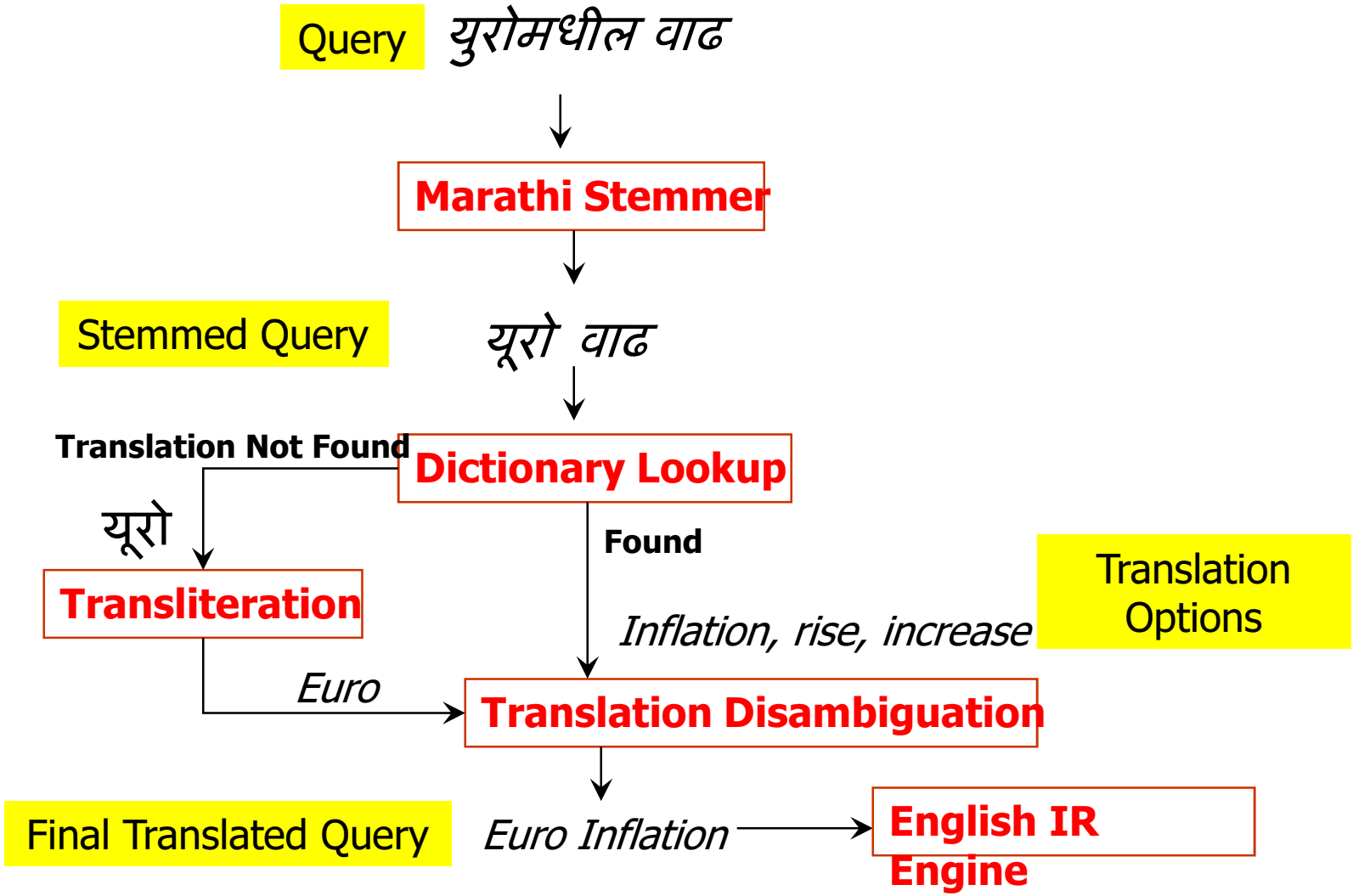
- Nucleus: The following can occur as nucleus:
  - All vowel sounds (monophthongs as well as diphthongs).
  - /m/, /n/ and /l/ in certain situations (for example, '*bottom*', '*apple*')

- Syllabic:
  - Both the onset and the coda are optional (as seen previously).
  - /j/ at the end of an onset (/pj/, /bj/, /tj/, /dj/, /kj/, /fj/, /vj/, /θj/, /sj/, /zj/, /hj/, /mj/, /nj/, /lj/, /spj/, /stj/, /skj/) must be followed by /uɪ/ or /ʊə/.
  - Long vowels and diphthongs are not followed by /ŋ/.
  - /ʊ/ is rare in syllable-initial position.
  - Stop + /w/ before /uɪ, ʊ, ʌ, aʊ/ are excluded.

Challenges in Machine Transliteration

- Lot of ambiguities at the grapheme level *esp.* while dealing with non-phonetic languages
  - Example: Devanagari letter *क* has multiple grapheme mappings in English *{ca, ka, qa, c, k, q, ck}*
- Presence of silent letters
  - Pneumonia – न्यूमोनिया
- Difference of scripts causes spelling variations *esp.* for loan words
  - रिलीस, रिलीज, जार्ज, जॉर्ज, बैंक, बँक

Introducing Transliteraion



Transliteration for OOV words

- Name searching (people, places, organizations) constitutes a large proportion of search
- Words of foreign origin in a language - *Loan Words*
  - Example: बस (bus), स्कूल (school)
- Such words not found in the dictionary are called *"Out Of Vocabulary (OOV) words"* in CLIR
- OOV words are usually automatically *"Transliterated"*

Machine Transliteration – The Problem

- Graphemes – Basic units of written language (English – 26 letters, Devanagari – 92 matraas)
- Definition

  "The process of automatically mapping an given grapheme sequence in source language to a valid grapheme sequence in the target language such that it preserves the pronunciation of the original source word"

Redefining Machine Transliteration

- Transliteration so far has been considered as an independent module used in Machine Translation, CLIR *etc.*
- In CLIR, important for term to be present in index
- In the above context, we redefine machine transliteration as

  "The process of automatically mapping an given grapheme sequence in source language to an index item in the target language index such that it preserves the pronunciation of the original source word"
- Pronunciation usually difficult to model – we only work with graphemes